DOCUMENT RESUME

ED 456 156 TM 033 219

AUTHOR Wang, Ning; Wiser, Randall F.; Newman, Larry S.
TITLE Use of the Rasch IRT Model in Standard Setting: An Item Mapping Method.
PUB DATE 2001-04-00
NOTE 36p.; Version of a paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Certification; *Cutting Scores; Interrater Reliability; *Item Response Theory; Licensing Examinations (Professions); Standards; *Test Items
IDENTIFIERS Angoff Methods; Mapping; *Rasch Model; *Standard Setting

ABSTRACT

This paper provides both logical and empirical evidence to justify the use of an item mapping method for establishing passing scores for multiple-choice licensure and certification examinations. After describing the item-mapping standard setting process, the paper discusses the theoretical basis and rationale for this newly developed method and presents the practical advantages of the item mapping method. Empirical evidence supporting the use of the item mapping method is provided by comparing the results from four standard setting studies for diverse licensure and certification examinations. The 4 cut scores, involving 79, 73, 43, and 53 items respectively, were conducted using both the item mapping and Angoff methods. Rating data from the four standard setting studies using each of the two methods were analyzed using item-by-rater random effects generalizability and dependability studies to examine which method yields higher inter-judge consistency. Results indicate that the item-mapping method produced higher inter-judge consistency and achieved greater rater agreement than the Angoff method. Another finding is that the item mapping method set consistently lower cut scores than the Angoff method. This result is consistent with other research findings that show unrealistically high ratings from the Angoff method or other test-centered standard setting methods that require judges to rate item by item. (Contains 3 tables, 3 figures, and 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 456 156

Use of the Rasch IRT Model in Standard Setting: An Item Mapping Method

Ning Wang

Ning Wang
Randall F. Wiser
Larry S. Newman

Assessment Systems, Inc.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033219

Running Head: An Item Mapping Standard Setting Method

An earlier version of this paper was presented at the 2001 annual meeting of the National Council on Measurement in Education, Seattle, WA. Corresponding concerning this article should be addressed to Ning Wang, Assessment Systems, Inc., Three Bala Plaza West, Suite 300, Bala Cynwyd, PA 19004.

BEST COPY AVAILABLE

Abstract

This paper provides both logical and empirical evidence to justify the use of an item mapping method for establishing passing scores for multiple-choice licensure and certification examinations. After describing the item-mapping standard setting process, the theoretical basis and rationale for this newly developed method are discussed. Also, the practical advantages of the item mapping method are presented. Empirical evidence supporting use of the item mapping method is provided by comparing results from four standard setting studies for diverse licensure and certification examinations. The four cut score studies were conducted using both the item mapping and Angoff methods. Rating data from the four standard setting studies, using each of the two methods, were analyzed using item by rater random effects generalizability and dependability studies to examine which method yields higher inter-judge consistency. Results indicated that the item mapping method produced higher inter-judge consistency and achieved greater rater-agreement than the Angoff method. Another finding is that the item mapping method sets consistently lower cut scores than the Angoff method. This result is consistent with other research findings that show unrealistically high ratings from the Angoff method or other test-centered standard setting methods that require judges to rate item by item.

Use of the Rasch IRT Model in Standard Setting: An Item Mapping Method

Introduction

Licensure and certification examinations are developed to determine whether candidates have adequate levels of knowledge and skills to perform competently in their professions. One of the most difficult steps in creating these examinations is determining a score that decides a passing level of competency. The pass/fail decision is high-stakes, both for candidates, and for the public that is served by candidates who pass. It is vital that the passing score reflect an appropriate standard of performance, a standard that differentiates between candidates who are sufficiently competent, and those who have not obtained the level of competence to be licensed. Defining and translating such a standard into a fair and legally defensible cut score requires a systematic application of a well-reasoned, reputable standard setting method (Berk, 1996; Mills, 1995; Norcini & Shea, 1997).

Among the standard setting methods, the Angoff (1971) procedure with various modifications is the most widely used for multiple-choice licensure and certification examinations (Plake, 1998). As part of the Angoff standard setting process, judges are asked to estimate the proportion (or percentage) of minimally competent candidates (MCCs) who will answer an item correctly. These item performance estimates are aggregated across items and averaged across judges to yield the recommended cut score. As noted (Chang, 1999; Kane, 1994; Impara & Plake, 1997), the adequacy of a judgmental standard-setting method depends on whether the judges adequately conceptualize the minimal competency of candidates, and whether judges accurately

estimate item difficulty based on their conceptualized minimal competency. A major criticism of the Angoff method is that judges' estimates of item difficulties for minimal competency are more likely to be inaccurate, and sometimes inconsistent and contradictory (Bejar, 1983; Goodwin, 1999; Mills & Melican, 1988; National Academy of Education, 1993; Reid, 1991; Shepard, 1995; Swanson, Dillon, & Ross, 1990). Studies found that judges are able to rank order items accurately in terms of item difficulty, but they are not particularly accurate in estimating item performance for target examinee groups (Impara & Plake, 1998; National Research Council, 1999; Shepard, 1995). A fundamental flaw of the Angoff method is to require judges to perform the nearly impossible cognitive task of estimating the probability of MCCs answering each item in the pool correctly (Berk, 1996; NAE, 1993).

An item mapping method, which applies the Rasch IRT model to the standard setting process, has been used to remedy the cognitive deficiency in the Angoff method (McKinley, Newman, & Wiser, 1996). The Angoff method limits judges to each individual item while they make an immediate judgement of item performance for MCCs. In contrast, the item mapping method presents a global picture of all items and their difficulties in the form of a histogram chart (item map), which serves to guide and simplify the judges' process of decision-making during the cut score study. The item difficulties are estimated through application of the Rasch IRT model. Using the Rasch IRT model, item difficulty and candidate ability are placed on the same scale, and the difference between a candidate's ability and an item's difficulty determines the probability of a correct response (Grosse & Wright, 1986). The item mapping method

uses this feature of the Rasch model to help judges determine an ability level representing minimal competency in terms of item difficulties.

The purpose of the present paper is to provide both logical and empirical evidence to justify the use of the item mapping method for establishing cut scores for licensure and certification examinations. After describing the item-mapping standard setting process, the theoretical basis behind this newly developed method and the rationale for using the item mapping method are discussed. In addition, the practical advantages of the item mapping method are presented. Empirical evidence is provided by comparing results from four standard setting studies, that were conducted using both the item mapping and Angoff methods to set cut scores for the set of diverse licensure and certification examinations. In addition to reporting and discussing the cut scores of the standard setting studies, rating data from the standard setting studies were analyzed using item by rater random effects generalizability and dependability studies to address the following research question: Does the item mapping or Angoff method allow more consistent item performance estimates by judges (i.e., which method yields higher inter-judge consistency)?

The Item Mapping Standard Setting Process[1]

The item mapping procedure incorporates item performance in the standard setting process by graphically presenting item difficulties. In item mapping, all the items for a given examination are ordered in columns, with each column in the graph

[1] The description of the item mapping method in this section is limited to procedures that are specific to the item mapping method. General implementation procedures that are required in a standard setting process, such as reviewing test purposes and discussing characteristics of minimally competent candidates, are also applied to the item mapping method.

representing a different item difficulty. The columns of items are ordered from easy to hard on a histogram-type graph with very easy items toward the left end of the graph, and very hard items toward the right end of the graph. Item difficulties in the unit of logits are estimated through application of the Rasch IRT model. In order to present items on a metric familiar to the judges, logit difficulties are converted to scaled values using the following formula: scaled difficulty = (logit difficulty × 10) + 100. This scale usually ranges from 70 to 130. Figure 1 provides an example of an item map. In the example, the abscissa of the graph represents the re-scaled item difficulty. Any one column has items within two points of each other (e.g., the column labeled "80" has items with scaled difficulties ranging from 79 to 81). Within each column, items are displayed in order by item-id numbers and can be identified by color and symbol-coded test content areas. By marking item content areas of the items on the map, a representative sample of items within each content area can be rated in the standard setting process.

Insert Figure 1 Here

The goal of item mapping is to locate a column of items on the histogram where judges can reach consensus for the question: Does the MCC have at least a 50% chance of answering the items correctly?

The process works as follows. After an extensive discussion of the characteristics of MCCs, the item map histogram is presented to the standard setting committee (judges). An easy item is first selected from the graph by the standard setting facilitator. The judges are asked to review the item text, and independently decide whether a typical

MCC has a 50% chance of answering the item correctly. Since the selected item is very easy, most of judges would agree that the MCC is able to answer the item correctly, but at a greater chance than 50%. The facilitator now selects a more difficult item and repeats the process until a column of items is located where most items in the column have consensus from the judges that the probability of a correct response for the MCC is 0.50. The level of difficulty under this column of items represents the cut score.

Theoretical Basis and Rationale

The Rasch IRT model provides the theoretical basis for the item-mapping standard setting method. The basic premise underlying the Rasch IRT model is that observed item responses are governed by an unobservable ability variable, θ, and item difficulties. If a pool of items require similar skills, the probability of a candidate with an ability level of θ answering an item correctly can be modeled by the mathematical function (Hambleton & Swaminathan, 1985):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, \ldots, n$$

Where $P_i(\theta)$ is the probability that a randomly chosen candidate with ability θ answers item i correctly, and is an S-shaped curve with values between 0 and 1 over the proficiency scale; b_i is the i^{th} item difficulty level; n is the number of items in the test; e is a transcendental number whose value is 2.718….

When the assumptions underlying the Rasch model are tenable, item difficulties and candidate abilities can be estimated and placed on the same scale through Rasch scaling procedures. The Rasch model gives the exact probability of a correct response to

an item, given a candidate's ability and the item difficulty. A unique feature of the Rasch model is: if candidate ability equals item difficulty, then the probability of a correct answer is 50%. If a candidate's ability is higher than the item difficulty, then he/she will have a greater than 50% chance of answering the item correctly. If a candidate's ability is less than the item difficulty, then he/she will have a less than 50% chance of responding correctly to the item. By utilizing this distinguishing feature of the Rasch model, the item mapping method enables judges to establish the passing score at the point where the MCCs' ability level equals the item difficulty.

Researchers offered two reasons for judges being unable to accurately estimate item performance for MCCs in test-centered standard settings (Impara & Plake, 1997). Judges may have difficulty in conceptualizing MCCs, and judges may have difficulty in estimating percentage correct, even for clearly defined MCCs. The methods and strategies used in item mapping are considered effective resolutions to these problems.

Conceptualizing the minimal competency and translating the concept into an operational definition of MCCs is typically required in setting cut scores for licensure and certification examinations. An extensive discussion is usually involved to help judges develop a common understanding of minimal competency (Cizek, 1996; Mills, Melican, & Ahluwalia, 1991). Even so, to conceptualize a group of hypothetical candidates is still considered to be a very difficult task for judges in test-centered standard settings (Impara & Plake, 1997). The item mapping method helps judges with the conceptualization of the MCC by asking them to consider a real and typical MCC in a training program or class. Then, the question for judges in item mapping is whether this real candidate has at least a 50% chance of answering an item correctly (i.e., a yes or no answer), rather than

providing an estimate of item performance for a group of MCCs. This strategy is considered to be beneficial to the judges' thinking process, and helpful in reducing the cognitive complexity in conceptualizing the MCC. Research reported that judges found conceptualizing a single, typical MCC to be easier and more realistic than imagining a group of hypothetical MCCs (Impara & Plake, 1997).

The item mapping method also helps reduce the cognitive complexity in estimating item performance for the MCC in several ways. The ease of considering a real and typical MCC helps judgements on the probability of the MCC answering an item correctly. Studies show that teachers are more accurate in estimating the performance of individual students in their class, than estimating the performance for the total group of students collectively (Hoge & Coladarci, 1989; Impara & Plake, 1997). Also, asking judges to consider a given probability (i.e., 50%) for the MCC answering an item correctly (i.e., providing a yes or no answer) involves less cognitive process than having the judges estimate the absolute probability for the MCC answering the item correctly.

The global picture of item difficulties (i.e., the item map) provides judges an intuitive and relative standing of items in terms of item difficulties. This picture serves as an external benchmark in guiding judges' decision-making processes. Studies found that standard setting results were more reproducible when item performance data were provided (Norcini, Shea, & Kanya, 1988; Swanson, Dillon, & Ross, 1990). The item mapping method provides not only item performance data, but also a relative relationship of items in a simple histogram. The modern theory of cognitive science indicates that for human information-processing, visual representation is less cognitively demanding than other modes of representation, such as symbolic and verbal (Simon, 1989). Therefore,

the strategy of visually presenting item difficulties in a global item map simplifies the judging process in a standard setting, and helps judges provide accurate and consistent estimates of the MCC's item performance.

Practical Advantages

Practical feasibility is an important criterion to consider in the selection of a standard setting method (Kane, 1995). In addition to having a strong theoretical basis, the item mapping method benefits from a number of practical advantages.

Portrait of the Minimally Competent Candidate's Ability. Item mapping is a graphical method of determining a minimal level of competency. A map of item difficulties is presented to judges during the standard setting process. The unique feature of the Rasch model, that candidates with a 50% chance of answering an item correctly have an ability level equal to the item difficulty, enables the determination of an ability level for the MCC from the item difficulty map. Once a column of items is identified where the MCC has a 50% chance of answering the items correctly (i.e., the cut score point), the standard setting process produces a complete picture of the MCC's ability in terms of the probabilities of correctly answering any item on the item map. Figure 2 provides a probability ruler used by the item mapping method. With the probability ruler aligned so that the 50% marker is centered over the column selected to be the cut, the probabilities of a correct response for all items for the MCC can be identified. By using these probabilities, the MCC's ability is portrayed for any position on the item map. This complete picture of the MCC's ability is very helpful in understanding the MCC's test

performance, and also very useful in determining whether more iteration of item ratings is needed to adjust the final cut score during the standard setting process.

Insert Figure 2 Here

Immediate Prediction of Passing Rates. Near the end of the standard setting study, a histogram representing the distribution of candidate performance from the entire group of test takers is presented. A line can be drawn at the ability level represented by the proposed cut score, yielding an immediate pass rate projection based on the cut score. This gives judges feedback on the reasonableness of the proposed cut score, given the distribution of candidate abilities. Figure 3 provides an example of an ability distribution histogram. In this example, if a column of items with a scaled difficulty of 115 is selected as the cut score, the predicted pass rate is approximately 84%.

Insert Figure 3 Here

Less Time Consuming. Another advantage of the item mapping method is that only a representative sample of items drawn from the item pool needs to be rated by judges, which reduces the time spent on the standard setting process. Instead of rating each item one by one as in the Angoff method, the item mapping method allows more time to be spent discussing those items where consensus is not easily reached. To rate fewer items also reduces the cost of a standard setting study. Experience indicates that for an item pool consisting of 300 items, two days are usually needed to rate all items

using the Angoff method, whereas only one day is needed to set a cut score using item mapping.

Consensus and More Discussion. Since the goal of a standard setting study is to reach a final cut point, most test-centered standard setting methods take an average of ratings when judges do not agree. Item mapping works by having consensus from the entire committee on where the cut should be established. Because of this emphasis on consensus, item mapping usually generates more discussion than the Angoff method. One hundred percent consensus may not occur, but the discussion aimed at forcing consensus is the process that ultimately allows the cut to be established. More discussion serves to eliminate unnecessary variations among judges, and to maintain attention during the cut score study.

Four Standard Setting Studies

Data Sources

Rating data are analyzed in this study from four recently conducted standard setting studies, where both item mapping and Angoff methods were used. The cut score studies were performed on four diverse multiple-choice professional licensure or certification examinations. The first three standard setting studies were conducted from August to November, 1999, and the last one was conducted in December, 2000.

Six judges participated in the first and second cut score studies, where 79 items for the first exam and 73 items for the second exam were rated by the same judges using both item mapping and Angoff methods. Thirteen judges participated in the third cut score study where 43 items were rated by the same judges using both the Angoff and item

mapping methods. For the fourth standard setting study, the same eight judges rated 53 items using both Angoff and item mapping methods.

All judges are subject matter experts in their professions. They are selected based on two criteria, which are for geographical representation of the test taker population and for representation of the profession.

To minimize the order effect of the standard setting methods, the item mapping method was used first in the first two studies, while the Angoff method was used first in the third and fourth studies.

Procedures and Data Analyses

Each of the cut score studies were conducted by an experienced test developer and/or psychometrician for both methods, using a standard procedure established for Assessment Systems, Inc.. All of the standard-setting studies began with an overview of the examination program and an introduction to standard setting in general. Then, the characteristics of MCCs were identified from an extensive discussion of the judges. Before judges used each method to rate items, the Angoff or item mapping method was presented to the judges and discussed in detail. For the Angoff method, judges were also given an opportunity to practice on a sample of test items prior to proceeding with actual ratings. For both the Angoff and item mapping methods, item statistics such as p-values and distractor analyses were shown to the judges after their initial ratings. Based on the item statistics and a discussion of their initial ratings with the group, judges were given the opportunity to change their ratings.

After each item mapping standard setting, the judges were asked to evaluate the cut score and determine whether it met their conceptual understanding of minimal competency. Additionally, the judges had the opportunity to express whether the predicted pass rate from the item mapping study met their expectation of the candidate population.

For the Angoff method, each judge's percentage rating on every item is recorded. These percentage ratings are aggregated across items and averaged across judges to yield the cut score for each examination.

When using the item mapping method, a value of "100" is recorded for an item if a judge agrees that a MCC has a 50% chance of answering the item correctly, otherwise, a "0" is recorded for the item.

To answer the question of which method yields greater inter-judge consistency, Rater by Item (r × i) random effects generalizability and decision studies are conducted for each of the eight sets (4 studies × 2 methods) of rating data. The magnitudes of variances due to sampling judges and items are compared for both methods across the four studies. The dependability coefficients from decision studies are also examined. These statistics provide information on the degree to which the item ratings for the MCCs are consistent across judges and across items (Brennan, 1992).

Results and Discussion

Table 1 provides estimates of the variance components from Rater by Item (r × i) random effects generalizability studies for each of the eight rating data sets (4 studies × 2 methods). The percentage distributions of variances for inter-item, inter-rater, and rater-

item interaction for each study across two methods are displayed in this table. For each standard setting method, distributions of variation due to each factor of the design are similar across the four studies. For the item mapping method, the variances due to items accounts for approximately 80% of the total variance across the four studies, whereas it only ranges from 40% to 50% for the Angoff method. While the variability due to raters accounts for a small amount of the total variance in both methods, a relatively larger percentage was found for the Angoff method (1.4% to 5%), as compared to the item mapping method (0% to 0.2%). Relatively larger proportions of variances are found due to the interaction between rater and item for the Angoff method (45% to 56%) than the item mapping method (16% to 21%). Given the single-faceted design of the analyses, this variance component is confounded with both measurement error and differential rating of items across raters. For the item mapping method, these results reveal that the variation due to item sampling dominates the total variance, while the variation due to sampling raters accounts for a negligible amount of the total variation. For the Angoff method, the variation due to measurement errors and differential rating of items across raters accounts for a large proportion of the total variance. The variation due to rater sampling accounts for a small amount of the total variance for the Angoff method, but still a larger amount than found for the item mapping method. These results indicate that judges provide more consistent ratings in the item mapping method than the Angoff method. So, the inter-judge consistency is higher in the item mapping method than in the Angoff method.

Insert Table 1 Here

Table 2 shows the dependability coefficients from Rater by Item (r × i) random effects dependability studies for the eight sets of rating data. The dependability coefficients yield information on the average consensus among judges in their item ratings; thus, serve as an index of rater agreement (Brennan, 1992; Brennan & Lockwood, 1980; Hurtz & Hertz, 1999). Results reveal that across the four studies, judges reach higher agreement in the item mapping method than in the Angoff method. All ratings from the item mapping method reach rater agreement higher than .95, whereas the rater agreements for the Angoff method range from .796 to .922. The degree of difference in rater agreement between the two methods varies from study to study. The second study shows the largest difference (.796 vs. .958), and the third study has the closest rater agreement (.922 vs. .985). The rater agreement index provides additional information on inter-judge consistency. Consistent with the findings from the variance components, the judges provide more consistent estimates of item performance using the item mapping method than the Angoff method.

Insert Table 2 Here

Table 3 provides the cut scores established by each method for each examination. The cut scores are reported in three scales: the number of items (raw cut score); the number of items in terms of percentage of the total items; and the converted item

difficulty scale used on the item map. It is noted that the item mapping method sets a lower cut score than the Angoff method. This is a consistent result for all four examinations. The magnitudes of the cut score differences vary from study to study, ranging from a difference of 9% to 19% in terms of the percentage of total items.

Insert Table 3 Here

Following the determination of the cut score in each of the item-mapping standard setting studies, judges were asked whether the cut score met their conceptual understanding of minimal competency. The majority of judges thought the cut scores were realistic, and met their conceptual understanding of the passing standard. From experience and observation using the Angoff method, results show that judges normally give item ratings between 70% and 80%; for easier items, ratings may rise to 90% or 95%, and for harder items, ratings may decrease to 50% or 60%. It rarely occurs that item ratings go below 50%. These rating ranges occurred not only in the standard settings reported in this study, but also in all other standard settings conducted by Assessment Systems, Inc. for licensure and certification examinations. As a result, Angoff cut scores are typically set between 70% and 80%. However, when the item mapping method is used, judges are more likely to be comfortable rating a certain number of difficult items, which a MCC has less than a 50% chance of answering correctly. Once a cut score is determined and the picture of the MCC's ability is portrayed, it appears that judges are comfortable with their relatively low ratings, since the final cut meets their conceptual understanding of minimal competency. The item

mapping method does broaden the range of cut scores relative to the Angoff method. If a broad range of cut scores exists in realty, then the accuracy (or validity) of the Angoff method would be in question. Research (Klein, 1984) has shown, that judges using the Angoff method based their performance estimates on average or above average candidates, rather than the MCC, resulting in ratings being unrealistically high. Also, overestimates of item performance for target candidate populations have been found in studies using the Angoff or other test-centered standard setting methods that require judges to conduct a similar task of rating item by item (Goodwin, 1999; Plake, 1998).

Summary and Conclusions

Licensure and certification testing provides a dual service. For the practitioners, it can provide employment opportunities. For the public, it ensures a certain level of safety. Neither service can be adequately provided without setting a justifiable and valid passing score. In current practice, the Angoff method is the most commonly used standard setting method (Kane, 1995; Plake 1998). Given the criticism of cognitive difficulty and inaccurate item performance ratings in the Angoff method, this paper advocates the use of the item mapping method for setting cut scores. Both logical and empirical evidence is provided to give legitimacy and support for the item mapping method.

The item mapping method uses the Rasch IRT model as a basis for standard setting. As part of the item-mapping standard setting process, judges are provided a histogram chart of items representing item difficulties, which are determined through application of the Rasch IRT model. The passing score is the difficulty/ability point on the chart where MCCs have a 50% chance of answering the items correctly. By requiring

judges to determine the ability level representing minimal competency in terms of item difficulties, the item mapping method minimizes judges' problems in accurately predicting item performance for MCCs.

Researchers found problems confronting judges in test-centered standard setting methods. One problem is conceptualizing a group of MCCs. Another problem is estimating item performance for the group of MCCs (Impara & Plake, 1997). The Item mapping method offers resolutions to these difficulties by reducing the cognitive complexity of the judges' decision-making processes. Asking judges to think of a real and typical MCC in a training program or class helps judges conceptualize the MCC, and simplifies the judges' thinking processes in determining the probability of an MCC answering an item correctly. Providing only a yes or no answer to the question, "Does the MCC have at least a 50% chance of answering an item correctly?", in item mapping is considered an easier task than estimating an absolute probability of a correct response in the Angoff method. The most important and unique feature of the item mapping method is that a global picture of item difficulties is presented to judges during the standard setting process. Visually presenting item difficulties on a global item map is an effective resolution alleviating the problem of judges being unable to provide accurate and consistent estimates of MCCs' item performance.

The item mapping method is appropriate for large-scale licensure and certification examination programs, where one single passing score is typically needed to make a high-stakes pass/fail decision. Given that multiple equivalent test forms are needed for such examinations to ensure test security, item mapping is applicable because of the resulting large number of test items used in setting cut scores. Since item mapping

reduces the time required for cut score studies, the method is beneficial when hundreds of items is used for a single cut score study. In addition, a relatively large number of test responses are usually collected from such examinations, allowing use of the Rasch IRT model, which subsequently provides data for creation of item maps. Even though the Angoff method with various modifications is the most prevalent standard setting method for such cut score studies, users and judges think that rating item by item is very time consuming and that rating fatigue may also be a concern in a long Angoff standard setting process.

The generation of a complete picture of the MCC's ability in terms of the probabilities of answering any item on the item map correctly, is a practical advantage of the item mapping method after the cut score is determined. This complete picture of the MCC's ability is very helpful in understanding the MCC's test performance, and in determining whether an initial cut score meets the judges' conceptual understanding of minimal competency. Sometimes, more iteration of item ratings may be needed to adjust the initial cut. A unique advantage given by examinee-centered standard setting methods is to provide an overall view of examinee performance. Although the item-mapping method appears to be a test-centered method, it also enjoys the advantage of providing judges with a global view of test performance, while allowing judges to work on detailed item-by-item determinations. This informative feature potentially enables the item mapping method to be more accurate than other test-centered approaches, such as the Angoff method, which only rely on item level judgments.

Another distinct advantage of the item mapping method is that the standard setting process aims at forcing consensus by generating extensive discussion among

judges, instead of taking an average of ratings like most test-centered standard setting methods. Other practical advantages of the item mapping method include savings on time, cost efficiency, and immediate feedback on the prediction of pass rates.

By analyzing data from four standard setting studies for a diverse set of professional licensure and certification examinations, this study also provides empirical evidence for better decision consistency and more justifiable results from the item mapping method than from the Angoff method. Statistical analyses of rating data from the four standard setting studies indicate that the item mapping method achieves higher rater agreement than the Angoff method. By examining the percentage distributions of the variance components from Rater by Item generalizability studies and the dependability coefficients from Rater by Item dependability studies, this study finds that the item mapping method allows for higher inter-judge consistency than the Angoff method.

Another finding of this study is that the item mapping method sets consistently lower cut scores than the Angoff method. Once a cut score is determined and the picture of the MCC's ability is portrayed, judges are comfortable with their relatively low ratings from the item mapping, because the final cut meets their conceptual understanding of minimal competency. This also enables the item mapping method to broaden the range of cut scores relative to the Angoff method. This is consistent with other research findings (Goodwin, 1999; Klein, 1984; Plake, 1998), that unrealistically high ratings may be obtained from the Angoff method, or other test-centered standard setting methods that require judges to rate item by item. Lack of an overall picture of item performance may

have contributed to less rating consistency and overestimation of item performance for the MCCs in the Angoff method.

As with any standard setting method, the item mapping method has its limitations. The use of this method requires previous statistical analysis of test data. All items and candidate abilities have to be estimated using the Rasch IRT model prior to the standard setting study. Also, the item mapping method requires more extensive preparation than the Angoff method (i.e., to generate the item map).

Researchers have recommended the use of an alternative version of the Angoff method which is a better choice than the traditional approach (Impara & Plake, 1997). Unlike the traditional Angoff method, the alternative approach asks judges simply to indicate whether a MCC will be able to answer each item correctly (yes/no method). The judgmental thinking used in this method is closer to the item mapping method than the traditional version of the Angoff method. To better recognize and appreciate the distinct advantages of the item mapping method, more research should be conducted to compare this yes/no version of the Angoff method with the item mapping method.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303-310.

Berk, R.A. (1996). Standard setting: the next generation. *Applied Measurement in Education*, *9*(3), 215-235.

Brennan, (1992). *Elements of generalizability theory*. Iowa City, IA: ACT.

Brennan, R. L. & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, *4*, 219-240.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education*, *12*(2), 151-166.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 13-21.

Goodwin, L. D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline candidates. *Applied Measurement in Education*, *12*(1), 13-28.

Grosse, M. E. & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Professions*, 9, 267-285.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory*. Boston, MA: Kluwer-Nijhoff.

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research*, *59*(3), 297-313.

Hurtz, G. M. & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, *59*(6), 885-897.

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*(4), 353-366.

Impara, J. C. & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, *35*(1), 69-81.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments*, Washington, DC.

Klein, L. W. (1984, April). *Practical Considerations in the Design of Standard Setting Studies in Health Occupations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

McKinley, D. W., Newman, L. S., & Wiser, R. F. (1996, April). *Using the Rasch model in the standard setting process*. Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY.

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purpose, procedures, and practices* (pp. 219-252). Lincoln, NE: Buros Institute of Mental Measurements.

Mills, C. N. & Melican, G. J. (1988). Estimating and adjusting cutoff scores: future of selected methods. *Applied Measurement in Education*, *1*, 261-275.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining Minimal Competence. *Educational Measurement: Issues and Practice*, *10*(2), 7-10.

National Academy of Education (1993). *Setting performance standards for student achievement*. Stanford, CA: Author.

National Research Council (1999). Setting reasonable and useful performance standards. In Perlligrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.), *Grading the nation's report card* (pp. 162-184).

Norcini, J. J. & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, *10*(1), 39-59.

Norcini, J. J. & Shea, J. A., & Kanya, D. T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement*, *25*, 57-65.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, *11*(1), 65-80.

Reid, J. B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice*, *10*(2), 11-14.

Shepard, L. A. (1995). Implications for standard setting of the national academy of education evaluation of the national assessment of educational progress achievement levels. *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments*, Washington, DC.

Simon, H. A. (1989). *Models of Thought* (Vol. II). New Haven: Yale University Press.

Swanson, D. B., Dillon, G. F., & Ross, L. E. (1990). Setting content-based standards for national board exams: initial research for the Comprehensive Part I Examinations. *Academic Medicine*, *65*, s17-s18.

Figure 1. Example Item Map

BEST COPY AVAILABLE

Figure 2. An Example of Probability Ruler

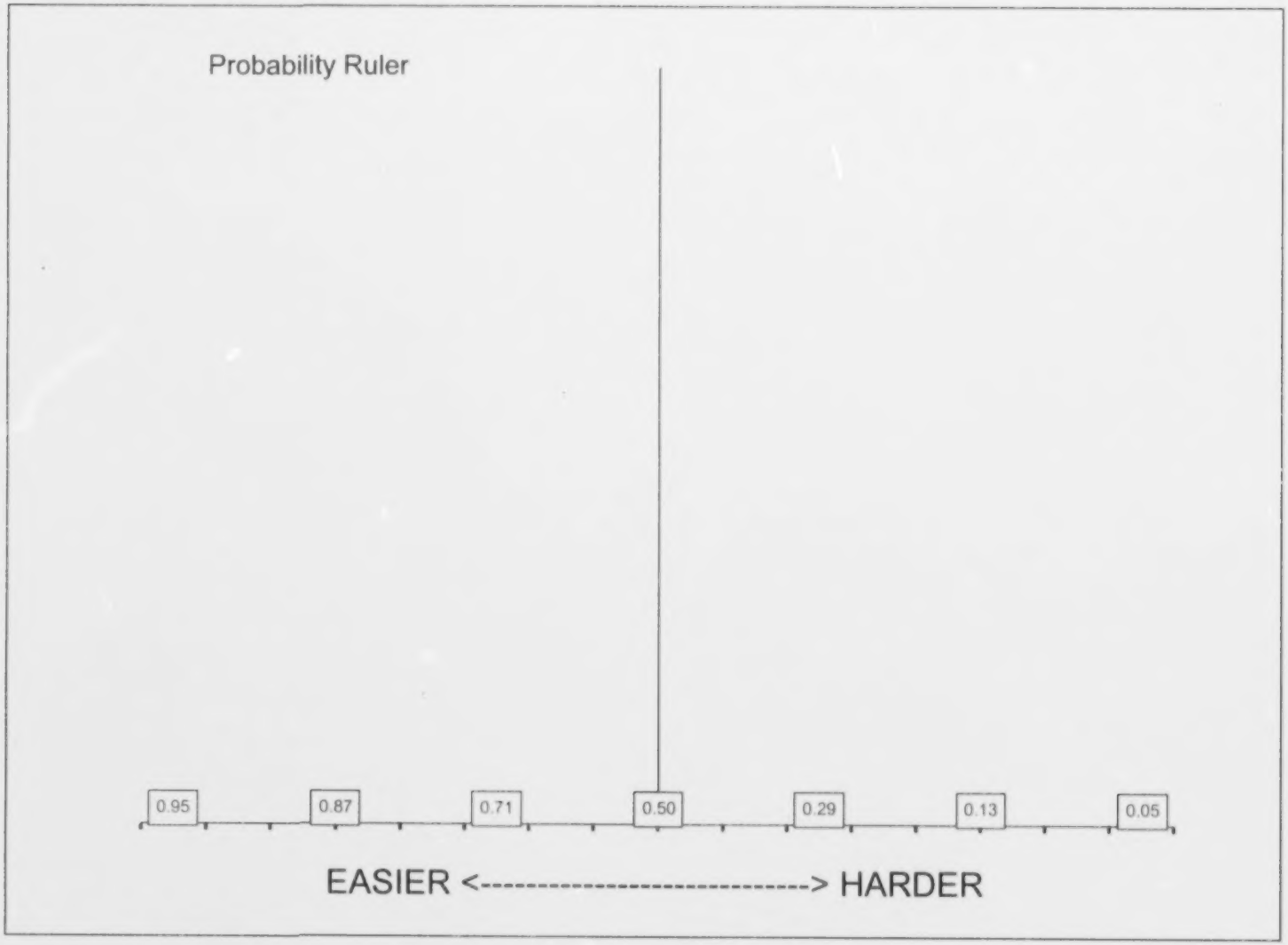

Figure 3. An Example of Candidate Ability Distribution

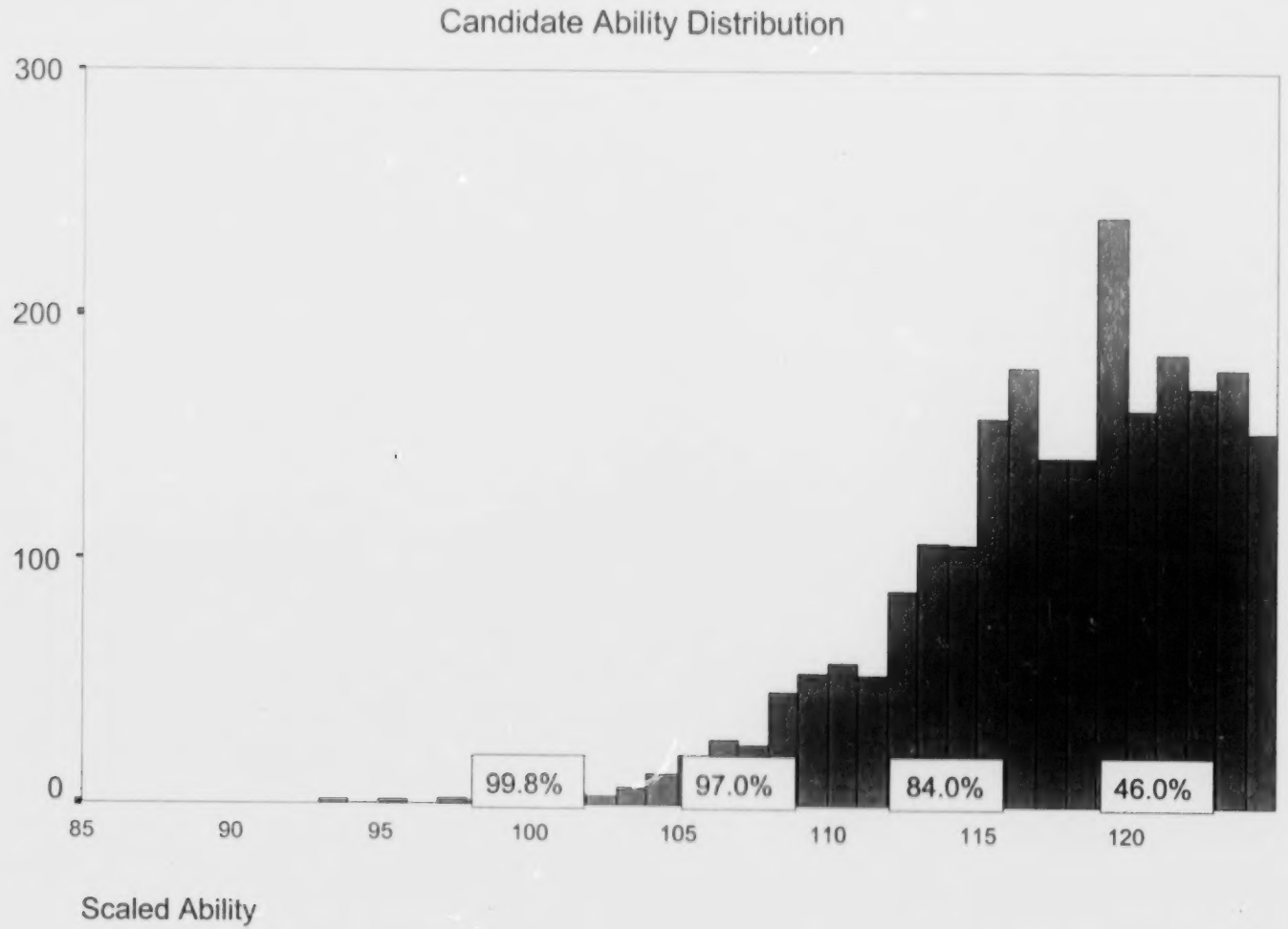

Table 1

Percentage Distributions of Variance Estimates for Item × Judge G-Studies

	1st study 6 judges & 73 items		2nd study 6 judges & 79 items		3rd study 13 judges & 43 items		4th study 8 judges and 53 items	
	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map
ITEM	49.6	81.9	39.5	79.1	47.6	83.5	50.2	79.0
JUDGE	4.5	0.2	4.9	0.1	1.4	0.1	4.9	0.0
ITEM × JUDGES	45.9	17.9	55.7	20.7	51.0	16.3	44.9	21.0

Table 2

Dependability Coefficients (Rater Agreement) for Item × Judge D-Studies

	1st study 6 judges & 73 items		2nd study 6 judges & 79 items		3rd study 13 judges & 43 items		4th study 8 judges and 53 items	
	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map
ϕ	0.856	0.964	0.796	0.958	0.922	0.985	0.890	0.968

Table 3

Raw Cut Scores and the Logit Cuts Established by Each Standard Setting Method

	1st study 6 judges & 73 items		2nd study 6 judges & 79 items		3rd study 13 judges & 43 items		4th study 8 judges and 53 items	
	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map	Angoff	Item Map
Raw Cut Score (%)*	49 (67%)	36 (49%)	53 (67%)	38 (48%)	29 (68%)	22 (51%)	43 (81%)	38 (72%)
Scaled Logit Cut	109	101	110	101	111	101	118	108

* A number in parenthesis represents the number of items in percentage.